Automatic Learning of Probabilistic Causal Models

Brenda M. Ng
(925) 422-4553
ng30@llnl.gov

Probabilistic causal models (PCMs) represent non-deterministic cause-and-effect relationships between stochastic entities. These models are crucial for causal reasoning, which ultimately drives informed decision-making, allowing us to achieve situational understanding about complex processes and potential threats that are central to LLNL's mission. Good examples are how one might account for the uncertainties and variability in the preparations for NIF ignition experiments or decide which countermeasure is the most effective against an epidemic. Through PCMs, we can infer these types of mission-critical intelligence. PCMs offer a more informative model than semantic graphs (our current technology), which merely provide associative information between entities and lack the infrastructure necessary for efficient reasoning under uncertainty. Automatic learning of PCMs can reveal hidden patterns and entities, thus providing template models (*e.g.*, Bayesian networks (BNs)) for processes lacking well-defined physics and/or expert domain knowledge.

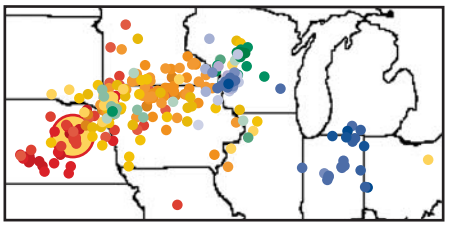

Figure 1. Map of simulated outbreak scenario. The infection source originated from a farm in Nebraska. The colored points denote the locations of infected farms. Red represents the earlier generations; blue represents the later generations. The first generation is defined as the set of farms infected by the source; the second generation consists of the farms infected by the first generation, and so on.

Project Goals

This effort focused on the automated learning of BNs as a means for uncovering causal information from observed data. The goal was to implement a toolbox of state-of-the-art algorithms for learning BNs, which encapsulate independence assumptions and stochastic dynamics among process variables.

Relevance to LLNL Mission

This work is highly relevant to furthering LLNL's missions related to Inference and Adversarial Modeling. The toolbox can 1) enhance pattern discovery, risk analysis, and predictive reasoning in any application related to modeling and decision-making under uncertainty, where the dynamics of agents are poorly understood, as in ignition experiments for NIF and disease propagation for the Biodefense Knowledge Center (BKC); 2) reduce time/labor and mitigate inconsistencies, uncertainties, and oversights associated with building models manually; and 3) be the launching pad for enhanced reasoning tools that integrate Bayesian modeling with general classes of computational expertise (*e.g.*, game-theoretic or agent-based) to improve optimal decision-making. Our results could be extended to decision-theoretic tools for a variety of applications including 1) predicting disease spread and assessing risk of nonintervention in an epidemic scenario; and 2) improving process modeling and experiments under uncertainty.

FY2008 Accomplishments and Results

A BN is a directed acyclic graph (DAG) that encodes a joint probability distribution over a set of random variables, in a factored manner that makes explicit use of the conditional independencies between variables. Given data from an unknown process, the goal is to perform BN learning on the data to estimate a model of the underlying process. Focusing on BN technology, our accomplishments include 1) a survey report of advanced machine learning methods; 2) implementations of learning methods; and 3) demonstrations on available programmatic data.

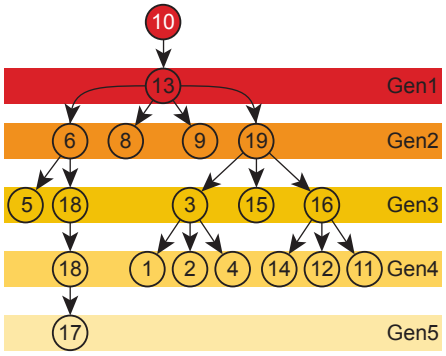

Figure 2. Trace tree of a particular outbreak scenario. Traces (*e.g.*, generation information) offer valuable information about the propagation of the disease outbreak.

BN learning involves uncovering the graphical structure, as well as the parameters that quantify probabilistic influences between variables. Structural learning is generally hard, because the number of possible structures (*i.e.*, DAGs) grows super-exponentially in the number of attributes. Evaluating all structures is intractable by traditional means.

Our toolbox implements two state-of-the-art algorithms. Both assume pre-specification of the number of variables and the cardinality of possible states. The first computes probabilities for all potential edges simultaneously, faster than any previous Markov Chain Monte Carlo (MCMC)-based samplers, and is applicable for edge discovery in networks. The second speeds up MCMC sampling of DAG structures, by applying dynamic programming for updating the proposal distribution to judiciously

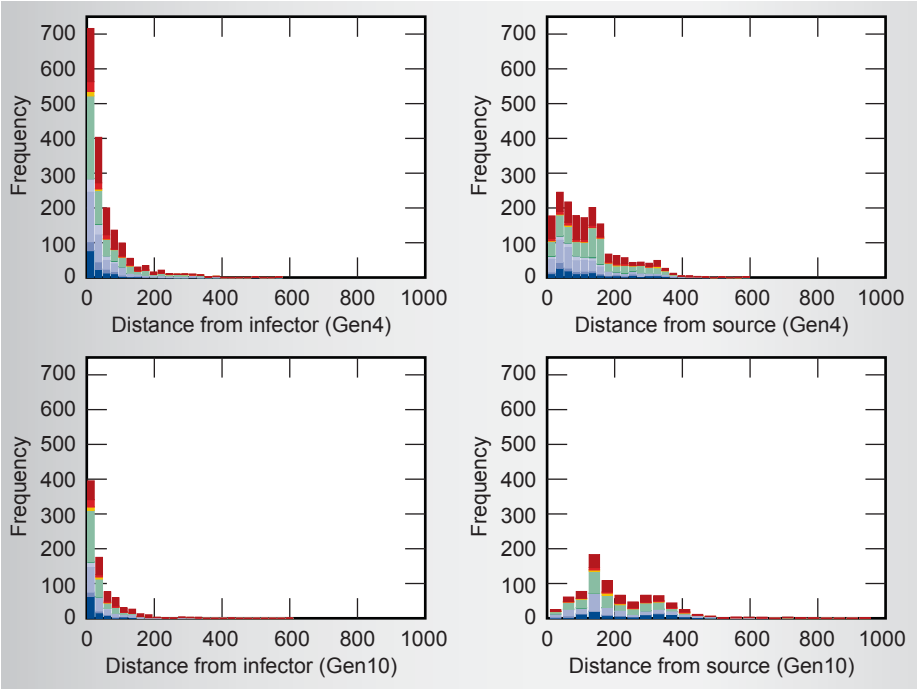

Figure 3. Stacked histograms comparing two generations' frequency counts in terms of distances from the infector (*i.e.*, the infecting farm from the previous generation) and the source (*i.e.*, the farm from which the outbreak originated). The constituent colors on each bar reflect the different farm types that make up the generation. From an earlier to a later generation, the outbreak spreads to similar farm types but is gradually moving away from the source.

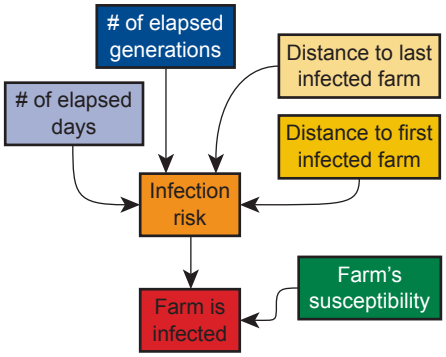

Figure 4. Bayesian network constructed from the MESA data. Using this, we want to assess whether a farm is infected as a function of the observable variables in the graph.

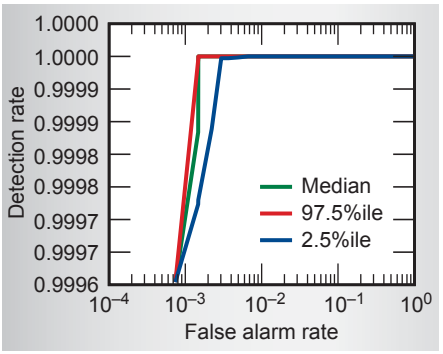

Figure 5. Receiver-operator characteristics (ROC) curves derived from five-fold cross-validation experiments. The ROC curve shows the tradeoff between the true positive rate (*i.e.*, the fraction of correctly identified infected farms) and the false positive rate (*i.e.*, the fraction of non-infected farms incorrectly identified as infected). The ROC curves show near-optimal performance with high detection rates and low false alarm rates.

guide the exploration through the space of DAGs.

The implemented algorithms were applied to simulated data sets from two sources: 1) energetics/diagnostics data from NIF; and 2) epidemiological data from BKC's Multiscale Epidemiological/Economic Simulation and Analysis (MESA) project.

We present findings here only for the MESA data. The data was derived from 401 independent simulations (resulting in 15667 infected farms) of a hypothetical, non-intentional foot-and-mouth disease outbreak. Figure 1 shows the extent of the outbreak from one simulated scenario. Each simulation contains the outbreak history that specifies the infection source, which farm(s) were infected next as time progressed, along with the farms' spatial locations and types. For each simulation, we performed tracing (Fig. 2) and derived generations linking the infected farms. In characterizing the data (Fig. 3), we have found that trace information, along with farm types and distances, are key attributes in detecting disease propagation. Putting this together, we constructed a BN (Fig. 4) consisting of these and other variables. Our results are summarized in Fig. 5.

Related References

1. Koivisto, M., "Advances in Exact Bayesian Structure Discovery in Bayesian Networks," *Proc. 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

2. Wood, F., T. L. Griffiths, and Z. Ghahramani, "A Non-Parametric Bayesian Method for Inferring Hidden Causes," *Proc. 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

3. Mansinghka, V., C. Kemp, J. Tenenbaum, and T. Griffiths, "Structured Priors for Structure Learning," *Proc. 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

4. Eaton, D., and K. Murphy, "Bayesian Structure Learning Using Dynamic Programming and MCMC," *Proc. 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

5. Fox, E. B., E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An HDP-HMM for Systems with State Persistence," *Proc. 25th International Conference on Machine Learning*, 2008.

FY2009 Proposed Work

FY2009 focus is on nonparametric modeling methods (*e.g.*, structured priors, infinite-state Hidden Markov Models, and hidden variable discovery), which will enable the discovery of newly active states and/or entities involved in a process. This work will produce powerful tools for generating more flexible and realistic models, thus enhancing complex system modeling, inference, and decision-making.